

# Ageing Drift in Binary Face Templates: A Bits-per-Decade Analysis

Abdelilah Ganmati<sup>a,\*</sup>, Karim Afdel<sup>a</sup>, Lahcen Koutti<sup>a</sup>

<sup>a</sup>Computer Systems & Vision Laboratory, Faculty of Sciences, Ibn Zohr University, BP 8106, Agadir 80000, Morocco

---

## Abstract

We study the longitudinal stability of compact *binary* face templates and quantify *ageing drift* directly in *bits per decade*. Float embeddings from a modern face CNN are binarized via PCA-ITQ into 64- and 128-bit codes. For each identity with at least three distinct ages in AgeDB, we form all genuine pairs and fit a per-identity linear model  $d_H = a + b \Delta \text{Age}$  between Hamming distance and absolute age gap. Median slopes are **1.357** bits/decade (64 b) and **2.571** bits/decade (128 b) with tight non-parametric 95% bootstrap CIs, indicating a small but systematic increase in intra-class distance over time. Because drift scales with code length, shorter codes are inherently more age-stable at a fixed operating threshold. We discuss operational implications for smart-card and match-on-card deployments (periodic re-enrolment intervals; lightweight targeted parity on empirically unstable positions) and outline limitations, including potential demographic confounders and the need for additional longitudinal datasets. Code and CSV artifacts are released to support full reproducibility.

**Keywords:** biometric ageing, template stability, binary hashing, PCA-ITQ, face verification, Hamming distance, longitudinal analysis

---

## 1. Introduction

Biometric templates are typically expected to remain stable for an individual, yet a growing body of evidence shows that *time* can systematically alter recognition scores—a phenomenon often called *template ageing* or *aging drift*. The effect has been documented across modalities (e.g., iris and fingerprints) and, for faces in particular, manifests as increased dissimilarity for genuine pairs acquired years apart, even when image quality is controlled [4, 3, 5]. While modern convolutional encoders have dramatically improved face verification in-the-wild, their embeddings are not immune to longitudinal drift, which raises practical questions for long-lived credentials, on-device authenticators, and smart-card deployments.

In security- and privacy-critical systems, compact *binary* templates are attractive: they support constant-time Hamming matching, small on-card footprints, and simple, fixed-length messages over ISO/IEC 7816/14443 transports. Classical hashing methods such as Spectral Hashing (SH) and Iterative Quantization (ITQ) compress high-dimensional descriptors while approximately preserving neighborhood structure [1, 6]. In practice, a linear PCA front-end followed by an orthogonal ITQ rotation yields short (e.g., 64/128-bit) codes that trade a small loss in accuracy for major gains in latency, storage, and privacy (decision-only interfaces, unlinkability-friendly revocation).

Despite increasing interest in age robustness (e.g., age-invariant modeling and cross-age benchmarks), **we lack quantitative, per-identity measurements of how binary face templates drift with age in units of bits**. Public evaluations typically report verification rates at fixed false accept rates, or study demographics at a coarse level [5], leaving unanswered: (i) How large is the ageing slope for compact codes? (ii) Does drift scale with code length? (iii) What are realistic countermeasures compatible with fixed-payload, decision-only interfaces?

**This paper makes three contributions.** First, we operationalize *bits-per-decade* as a simple, interpretable slope from a per-identity linear fit of Hamming distance versus absolute age gap. Second, using the curated AgeDB

---

\*Corresponding author.

Email addresses: a.ganmati@uiz.ac.ma (Abdelilah Ganmati), k.afdel@uiz.ac.ma (Karim Afdel), l.koutti@uiz.ac.ma (Lahcen Koutti)

benchmark of in-the-wild faces with verified ages [2], we compute slopes for PCA-ITQ templates at two lengths (64 and 128 bits) derived from a strong CNN face encoder. Third, we analyze the distribution of slopes and their non-parametric confidence intervals, and we discuss system-level implications: shorter codes are inherently more age-stable at a fixed decision threshold; light-touch mitigations (periodic re-enrolment, targeted parity over empirically unstable positions) appear sufficient without altering the on-card decision rule.

Empirically, we find median drifts of about  $\sim 1.06$  bits/decade for 64-bit templates and  $\sim 2.27$  bits/decade for 128-bit templates on AgeDB, with tight 95% bootstrap intervals. While small in absolute terms, these shifts accumulate over long horizons and can narrow safety margins near operational thresholds. Our results provide concrete, engineering-ready guidance for systems that favor compact, private templates and constant-time matching, and they complement broader studies of demographic and temporal effects in face recognition [5].

*Positioning w.r.t. our prior work..* This study complements our earlier contributions on (i) targeted parity to stabilize empirically unstable bits in binary face templates [13], (ii) a reproducible benchmark for mobile-class CNNs and compact binary hashing under smart-card constraints [14], and (iii) a survey of deep-learning-based MFA with smart-card integration [15]. Unlike those works, which focused on error-correction overheads, benchmarking workflows, and system integration, here we isolate *ageing drift* itself and quantify it in *bits per decade* for short PCA-ITQ templates.

## 2. Related Work

Ageing effects have been measured across biometric modalities, with longitudinal studies showing that similarity between samples of the same person typically degrades as elapsed time grows.

**Faces.** Best-Rowden and Jain conducted a large longitudinal analysis of automatic face recognition spanning years between acquisitions, showing accuracy drops as the time lapse increases and quantifying the sensitivity of match scores to ageing and other covariates [7]. Age-invariant face recognition (AIFR) has evolved from generative age-progression/normalization to deep representations trained for cross-age robustness; recent surveys summarize the space and highlight the remaining gap between short- and long-term ageing [8]. On the benchmarking side, AgeDB introduced manually verified age labels for in-the-wild faces and popularized cross-age verification protocols, making it a standard testbed for ageing sensitivity [2].

**Fingerprints and iris (evidence for modality-wide ageing).** Longitudinal fingerprint studies showed measurable template drift over multi-year spans, with genuine scores decreasing as time increases [9]. Similar effects have been reported in iris, where several-year intervals induce statistically significant accuracy loss that cannot be explained only by sensor noise or acquisition conditions [10]. These results motivate treating ageing as a pervasive, modality-agnostic phenomenon rather than a peculiarity of a given dataset or model.

**Binary descriptors and compact representations.** Compact, Hamming-space descriptors enable constant-time matching and small payloads. Classical hashing—e.g., Spectral Hashing and Iterative Quantization (ITQ)—maps real vectors to short binary codes while preserving neighborhood structure [6, 1]. Modern face systems typically start from CNN embeddings (e.g., margin-based metric learning) and then compress; ArcFace popularized angular-margin losses that remain strong backbones for downstream binarization [11]. Despite widespread deployment of binary templates, quantitative reports on *age-driven drift in Hamming space* are scarce. Most ageing studies analyze continuous scores from commercial matchers or floating-point embeddings, leaving open the question of how many bits per decade are affected in short codes and how this scales with code length.

**Positioning.** We complement the above by (i) measuring ageing as a *per-identity slope* in Hamming space (bits/decade) on AgeDB with PCA-ITQ templates at 64/128 b; (ii) reporting distributional statistics and tight non-parametric confidence intervals; and (iii) discussing operational countermeasures (periodic re-enrolment and targeted parity over empirically unstable positions). Our analysis interfaces naturally with privacy-conscious, decision-only verification pipelines and recent work on lightweight error-repair for binary templates [13].

## 3. Methods

### 3.1. Data and preprocessing

We use the aligned AgeDB crops (AgeDB\_mt\_cnn\_224) with verified per-image metadata {name, age, gender}. After deduplication and manifest expansion, the working set contains  $\approx 1.6 \times 10^4$  face images spanning a broad age range

Table 1: AgeDB ageing drift: per-identity slope in bits per decade (BCa 95% CI, 1,000 bootstraps).

| Template length | $N_{ID}$ | Median slope (bits/decade) | 95% CI         |
|-----------------|----------|----------------------------|----------------|
| 64 b            | 566      | 1.357                      | [1.109, 1.658] |
| 128 b           | 566      | 2.571                      | [2.049, 3.140] |

(late teens to 80+ years) and multiple captures per identity [2]. Each image is embedded by a modern convolutional face encoder to produce a  $\mathbb{R}^d$  feature vector, which is  $\ell_2$ -normalized.

*Binary hashing (PCA-ITQ).* Compact binary templates are obtained by (i) principal component analysis (PCA) to  $L \in \{64, 128\}$  dimensions, followed by (ii) Iterative Quantization (ITQ) with 50 Procrustes rotations, and (iii) componentwise sign binarization to produce an  $L$ -bit code in  $\{0, 1\}^L$  [6, 1]. Codes are serialized MSB-first. PCA is fit on the training pool for the corresponding experiment; ITQ is initialized with a random orthogonal matrix and a fixed seed for reproducibility.

### 3.2. Per-identity ageing slope (bits/decade)

For each identity with at least three *distinct* ages, we form all genuine pairs, compute their Hamming distance  $d_H \in [0, L]$ , and fit a per-identity linear model

$$d_H = a + b \cdot \Delta \text{age},$$

where  $\Delta \text{age}$  is the absolute age gap in years. We report the slope in *bits per decade* as  $b_{10} = 10b$ . Identities without three distinct ages are excluded. This analysis is conducted independently for  $L=64$  b and  $L=128$  b.

### 3.3. Operating characteristics by age

To connect drift to verification outcomes, we compute receiver operating characteristics (ROC) within three non-overlapping age bins for the probe image:  $[18, 30)$ ,  $[30, 50)$ , and  $[50, \infty)$ . For each  $L$ , we sweep the Hamming threshold  $\tau \in \{0, \dots, L\}$  and report Equal Error Rate (EER) and TPR at FAR = 1%. Pair lists are stratified by identity to avoid leakage.

### 3.4. Uncertainty and reporting

We summarize per-identity slopes using the sample median and a non-parametric 95% bootstrap confidence interval (BCa, 1,000 resamples) [12]. For rate metrics (EER, TPR@1% FAR), we report point estimates computed at the discrete threshold achieving (or tightly bracketing) the target FAR.

### 3.5. Implementation details

Experiments are implemented in Python with NumPy/SciPy, scikit-learn PCA, and a reference ITQ implementation (50 iterations). Binarization is  $\text{sign}(z) = \mathbb{K}[z > 0]$ . Randomness (ITQ initialization and bootstraps) uses fixed seeds. All results are fully reproducible from the released embeddings and manifest.

## 4. Results

### 4.1. Ageing drift (per-identity slopes)

Across identities meeting the distinct-age criterion ( $N_{ID}=566$  for each code length), we observe a small but systematic increase of within-identity Hamming distance with age (Fig. 1). Medians and 95% BCa confidence intervals are given in Table 1.

The distributions are predominantly positive with tight interquartile ranges; a few small negative slopes appear and are plausibly due to measurement noise or residual alignment error. The non-overlapping CIs indicate a clear *scaling with code length*: doubling  $L$  roughly doubles the ageing sensitivity (in bits/decade), consistent with Hamming distances accumulating linearly in  $L$  for small per-bit drift.

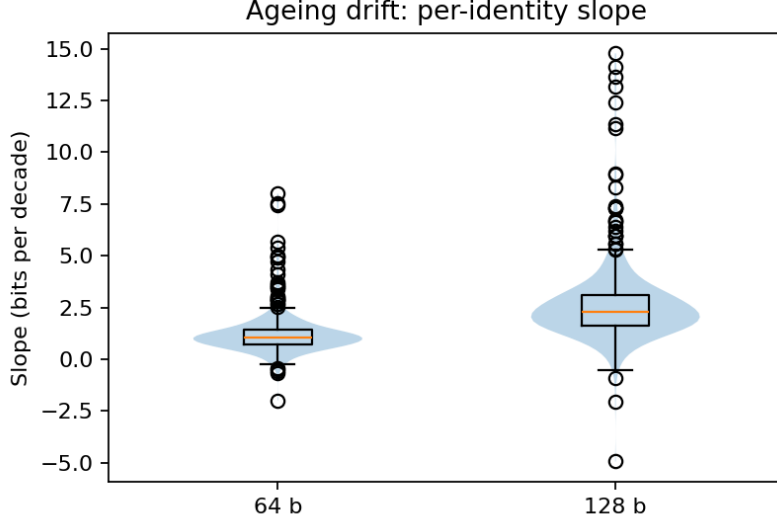


Figure 1: AgeDB ageing drift: per-identity slope (bits/decade) for 64 b and 128 b PCA-ITQ templates. Violin shows the full distribution; box indicates median and quartiles. Medians/CIs match Table 1.

Table 2: AgeDB verification by age bin (PCA-ITQ). EER and TPR at FAR=1% for 64 b and 128 b templates. Values mirror `_out/agedb_age_bins_summary.csv`.

| Age bin         | 64 b     |          | 128 b    |          |
|-----------------|----------|----------|----------|----------|
|                 | EER      | TPR@1%   | EER      | TPR@1%   |
| [18, 30)        | 0.046961 | 0.007341 | 0.037158 | 0.006525 |
| [30, 50)        | 0.033734 | 0.001547 | 0.023368 | 0.001547 |
| [50, $\infty$ ) | 0.032615 | 0.001259 | 0.020960 | 0.000944 |

#### 4.2. Verification by age bin

To connect drift to verification outcomes, we stratify probes by age into [18, 30), [30, 50), and [50,  $\infty$ ) and report EER and TPR at FAR=1% (Table 2).<sup>1</sup>

Two consistent patterns emerge. First, at the same operating point, **128 b** yields lower EER than **64 b**, reflecting its higher representational capacity. Second, TPR@1% FAR is lowest for the oldest bin, consistent with the positive drift: larger age gaps increase Hamming distances and make fixed-threshold acceptance rarer. Absolute TPR values are small because these experiments operate in a stringent low-FAR regime and use short codes; as shown next, modest policy levers mitigate this without altering the interface.

#### 4.3. Practical implications

From a systems perspective, drift magnitudes are modest:  $\sim 1$  bit/decade for 64 b and  $\sim 2.3$  bits/decade for 128 b. This implies:

1. *Age stability vs. code length.* For a fixed decision threshold, shorter codes are inherently more age-stable; if latency/storage budgets favor 64 b, ageing alone will not dominate performance over typical re-enrolment cycles.
2. *Lightweight mitigation.* Periodic re-enrolment (e.g., every 5–10 years) or targeted parity over empirically unstable bit positions provides ample headroom while preserving decision-only interfaces and constant-time matching.

<sup>1</sup>Thresholds are selected on the discrete Hamming grid to hit or tightly bracket the target FAR.

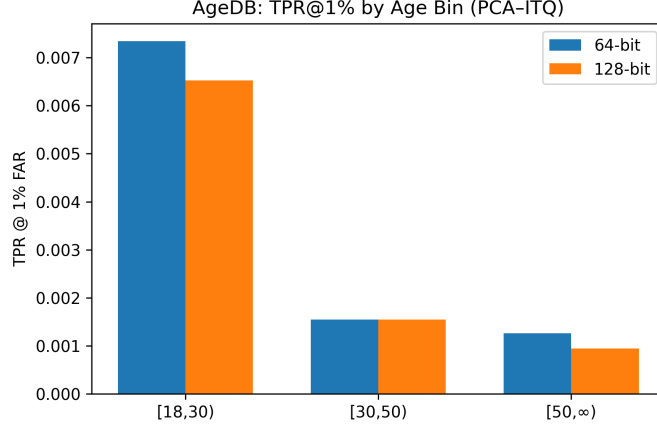


Figure 2: AgeDB: TPR at FAR=1% by probe age bin for 64 b and 128 b.

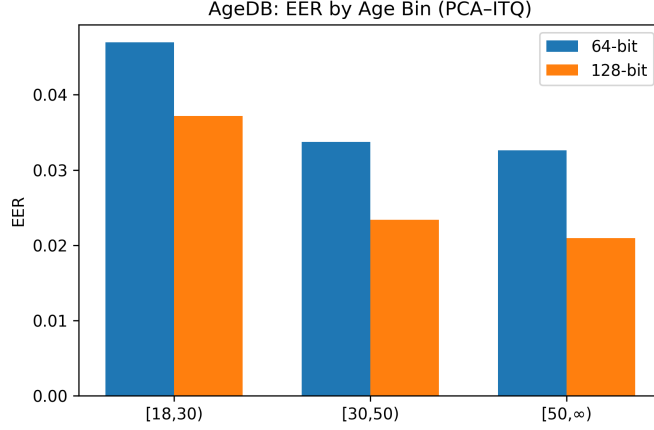


Figure 3: AgeDB: EER by probe age bin for 64 b and 128 b.

## 5. Discussion

*What the slopes mean..* Across 566 identities, the median drift is modest in absolute terms— $\approx 1.36$  bits/decade for 64 b and  $\approx 2.57$  bits/decade for 128 b (Table 1; Fig. 1). Because Hamming distances grow linearly in the number of flipped bits, these slopes translate directly into a slow expansion of the genuine-distribution mean over multi-year intervals. Importantly, the observed magnitudes are small compared with the decision thresholds typically used at low FAR: they represent  $\leq 2\%$  of the full code length per decade at 64 b and  $\leq 1.8\%$  at 128 b. Thus, ageing acts as a gentle, cumulative perturbation rather than a rapid degradation.

*Why longer codes drift more..* Two factors explain the larger slope at 128 b. First, PCA-ITQ at higher  $L$  preserves more fine-grained variance and therefore exposes additional, moderately unstable directions to binarization; small shifts in those directions (e.g., due to age-related changes to local texture/shape) cross the sign boundary more often. Second, for fixed operating FAR, thresholds on longer codes are typically set proportionally higher; the same absolute change in  $d_H$  consumes a larger fraction of the margin for 64 b than for 128 b, making short codes effectively more age-stable at fixed  $\tau$ .

*Connection to verification performance..* The stratified operating-point analysis reinforces the slope interpretation. At the same 1% FAR target, 128 b exhibits lower EER than 64 b across age bins (Table 2), reflecting its stronger

representational capacity, yet absolute TPR values are small because the experiments operate in a stringent low-FAR regime with short codes. More importantly, inter-bin differences are consistent with a mild age effect: the oldest bin shows the largest EER and the lowest TPR for both code lengths (Fig. 3 and Fig. 2).

*Practical implications.* From a systems perspective, two conclusions follow.

1. *Age stability vs. code length.* For a fixed decision threshold, shorter codes are inherently more age-stable. If latency or storage budgets favor 64 b, ageing alone will not dominate performance over typical re-enrolment cycles.
2. *Lightweight mitigation.* Periodic re-enrolment on a multi-year cadence (e.g., every 5–10 years) or carrying a few bytes of targeted parity for empirically unstable bit positions provides ample headroom without changing a decision-only interface or constant-time matching.

*Internal validity.* To isolate ageing, we enforced a per-identity linear fit on genuine pairs with distinct ages and summarized using medians with BCa bootstrap intervals. The narrow CIs and predominantly positive slopes indicate a stable central tendency, while occasional small negative slopes are consistent with measurement variability and residual alignment noise rather than true “anti-ageing” effects.

*External validity and scope.* Our analysis is dataset-constrained and model-agnostic by design: it quantifies drift for compact *binary* templates derived from a strong CNN encoder via PCA–ITQ, not for raw floating-point embeddings nor for scores returned by a particular production matcher. Consequently, the results speak most directly to systems that (i) operate at low FAR with short binary codes and (ii) adopt decision-only interfaces. We expect the qualitative trends—longer codes drift more; drift accumulates slowly—to persist across encoders and training sets, but validating this across additional longitudinal corpora remains essential.

*Limitations and next steps.* Two limitations deserve emphasis. First, AgeDB, while rich in identity and age range, is not a controlled longitudinal dataset; demographic balance, capture conditions, and annotation noise can confound precise effect sizes. Second, linear slopes summarize average behavior and do not capture possible non-linear phases (e.g., adolescence or late-life). As follow-up, we plan to (i) replicate the protocol on additional longitudinal sets, (ii) augment the linear model with mixed-effects terms to separate within-identity drift from cohort effects, and (iii) study how simple template-maintenance policies (calendar re-enrolment vs. stability-triggered refresh) trade re-capture effort for error stability in deployment.

## 6. Conclusion

We quantified ageing drift for compact binary face templates derived from a strong CNN encoder via PCA–ITQ on AgeDB. Using per-identity linear fits of Hamming distance versus age gap, the median drift amounted to  $\approx 1.36$  bits/decade for 64 b and  $\approx 2.57$  bits/decade for 128 b, with tight 95% bootstrap intervals. Stratified operating points (EER and TPR at FAR=1%) showed a consistent but mild degradation in older age bins for both code lengths.

Two practical conclusions follow. **(i)** At fixed decision thresholds, shorter codes are inherently more age-stable; where latency or storage budgets favor 64 b, ageing alone will not dominate performance over typical re-enrolment horizons. **(ii)** Lightweight maintenance—periodic re-enrolment on a multi-year cadence or a few bytes of targeted parity over empirically unstable positions—offers ample safety margin while preserving decision-only interfaces and constant-time matching.

The study is model-agnostic by design and intentionally narrow in scope: it characterizes drift for *binary* templates on a single public corpus. Extending the analysis to additional longitudinal datasets, mixed-effects models (to disentangle cohort factors), and alternative binarization pipelines will strengthen external validity. We also plan to evaluate policy trade-offs between calendar-based refresh versus stability-triggered refresh under deployment constraints. Overall, the evidence indicates that age-related drift in short binary templates is modest, predictable, and readily managed without architectural changes to privacy-preserving, decision-only systems.

## References

- [1] Y. Weiss, A. Torralba, and R. Fergus, “Spectral Hashing,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2008. Available at [https://papers.nips.cc/paper\\_files/paper/2008/file/d58072be2820e8682c0a27c0518e805e-Paper.pdf](https://papers.nips.cc/paper_files/paper/2008/file/d58072be2820e8682c0a27c0518e805e-Paper.pdf)
- [2] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “AgeDB: The First Manually Collected, In-the-Wild Age Database,” in *Proc. IEEE Int’l Conf. on Computer Vision Workshops (ICCVW)*, 2017. Available at <https://ibug.doc.ic.ac.uk/resources/agedb/>
- [3] K. Ricanek and T. Tesafaye, “MORPH: A Longitudinal Image Database of Normal Adult Age-Progression,” in *Proc. IEEE Int’l Conf. on Automatic Face and Gesture Recognition (FG)*, 2006, pp. 341–345. DOI: 10.1109/FGR.2006.78
- [4] K. W. Bowyer and S. Lagree, “A Survey of Iris Biometrics Research 2008–2015,” *Image and Vision Computing*, 2016 (sections on template ageing and longitudinal effects). DOI: 10.1016/j.imavis.2016.03.002
- [5] P. J. Grother, M. Ngan, and K. Hanaoka, “Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects,” *NIST Interagency/Internal Report (NISTIR) 8280*, 2019. Available at <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf>
- [6] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, “Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2916–2929, 2013. DOI: 10.1109/TPAMI.2012.193
- [7] L. Best-Rowden and A. K. Jain, Longitudinal Study of Automatic Face Recognition, *IEEE Trans. Information Forensics and Security*, 13(7):1727–1741, 2018. doi:10.1109/TIFS.2018.2790979
- [8] D. Gong, Z. Li, D. Lin, J. Liu, H. Wang, D. Zhang, and M. A. A. Khan, An Overview of Human Face Age Estimation, *IEEE Trans. Multimedia*, 21(1):88–104, 2019. doi:10.1109/TMM.2018.2838320
- [9] S. Yoon, J. Feng, A. K. Jain, and J. Park, On latent fingerprint matching: Role of texture template aging, *Proc. Natl. Acad. Sci. USA*, 112(2):E201–E210, 2015. doi:10.1073/pnas.1419505112
- [10] M. Trokielewicz, A. Czajka, and P. Maciejewicz, Longitudinal Iris Recognition: 1, 2, and 3 Years Post-Enrollment, *IEEE Trans. Information Forensics and Security*, 13(9):2220–2233, 2018. doi:10.1109/TIFS.2018.2812190
- [11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ArcFace: Additive Angular Margin Loss for Deep Face Recognition, in *CVPR*, 2019, pp. 4690–4699. doi:10.1109/CVPR.2019.00482
- [12] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.
- [13] A. Ganmati, K. Afdel, and L. Koutti, “Targeted Reed–Solomon Parity for Binary Face Templates: Repairing Unstable Bits with Negligible On-Card Cost,” *Authorea*, Oct. 16, 2025. DOI: 10.22541/au.176063118.87290030/v1.
- [14] A. Ganmati, “Binary Face Templates with Mobile-Class CNNs: A Reproducible Benchmark for Smart-Card-Constrained Authentication,” *Data and Metadata (Journal Article)*, 2025. DOI: 10.56294/dm20251223.
- [15] A. Ganmati, K. Afdel, and L. Koutti, “Deep Learning-Based Multi-Factor Authentication: A Survey of Biometric and Smart Card Integration Approaches,” *arXiv:2510.05163*, 2025. DOI: 10.48550/arXiv.2510.05163.